Audio Engineering Society

Convention Paper 7655

Presented at the 126th Convention
2009 May 7–10 Munich, Germany

Feature Selection vs. Feature Space Transformation in Automatic Music Genre Classification Tasks

Hanna Lukashevich[1]

[1] *Fraunhofer IDMT, Ilmenau, Ehrenbergstr. 31, 98693, Germany*

Correspondence should be addressed to Hanna Lukashevich (lkh@idmt.fraunhofer.de)

ABSTRACT

Automatic classification of music genres is an important task in music information retrieval research. Nearly all state-of-the-art music genre recognition systems start from the feature extraction block. The extracted acoustic features often could tend to be correlated or/and redundant, which can cause various difficulties in the classification stage. In this paper we present a comparative analysis on applying supervised Feature Selection and Feature Space Transformation algorithms to reduce the feature dimensionality. We discuss pros and cons of the methods and weigh the benefits of each one against the others.

1. INTRODUCTION

During recent years the scientific and commercial interest in automatic methods for music genre classification has tremendously increased. Stimulated by the ever-growing availability and size of digital music collections, automatic music genre classification has been identified as an increasingly important means to aid convenient exploration of large music catalogs. Due to the nearly unrestricted amount of musical data, real-world music genre classification systems have to be highly efficient and scalable.

At the first stage nearly all state-of-the-art music genre classification algorithms use acoustic features calculated in short time frames. Each feature is designed to correlate with one of the aspects of perceptual similarity, e.g. timbre, tempo, loudness, harmony etc. The distinct acoustical features are joined together into so called acoustical feature vector. While temporal changes in one feature often correspond to temporal changes in the other feature (for instance, timbre might be changing along with loudness), the individual dimensions of feature vectors can often be strongly correlated or/and redundant. Such raw feature vectors might cause various problems in the subsequent classification stage, for instance it could lead to singularity of covariance matrices while training Gaussian Mixture Models.

There are two principal ways to tackle this problem. The first approach is so called Feature Selection (FS). It's goal is to pick the feature dimensions which are most suitable for classification. In this work we apply a well established algorithm called Inertia Ratio Maximization (IRMFSP) with Feature Space Projection [1], [2]. Supervised by provided music genre labels, IRMFSP selects the feature dimensions providing maximal inertia ratio.

The second class of methods is Feature Space Transformation (FST). Here the original feature vectors are transformed (mapped) into the other, usually lower dimensional space. In Music Information Retrieval (MIR) the most commonly used FST method is Linear Discriminant Analysis (LDA) [3]. LDA is a supervised method which uses music genre labels to find the mapping with the maximum class separation. As LDA is a linear transform, it often does not provide an optimal solution. Among linear FST method we propose using their Nonlinear kernel counterpart, namely General Discriminant Analysis [4].

The choice of feature dimensionality reduction method also strongly depends on the applied classifier. In this paper we include four distinct classifiers in the evaluation. Two of them are commonly used and fairly well performing classifiers – Gaussian Mixture Models and Support Vector Machines – and need some optimization approximation on the training stage. Additionally, we compare them to two rather straight forward classifiers: Naive Bayes and k-nearest neighbor. In the evaluation we show how the above mentioned feature dimensionality reduction algorithms in the combination with the applied classifiers influence the achieved results in music genre classification framework.

2. FEATURE EXTRACTION

Nearly all state-of-the-art music similarity techniques use acoustic features calculated in short frames. Each feature is designed to correlate with one of the aspects of perceptual similarity, e.g. timbre, tempo, loudness, harmony etc. The usual practice of MIR algorithms is to use a compact representation of an audio signal derived in short-time signal snippets (frames). These representations (usually called “feature vectors” or “features”) are designed to correlate to some semantically meaningful properties of musical signal. Although distinct audio signals may possess audio properties, which can be captured only by signal-specific feature vectors, the MIR community has developed a set of state-of-the-art feature vectors well performing for the music genre classification.

In this paper we use a set of eight low-level features derived from adjacent 10 ms frames. This set includes a well established timbre descriptor - Mel-Frequency Cepstral Coefficients (MFCCs) [4], and several descriptors capturing rhythm loudness or frequency related information, proposed within the MPEG-7 standard [5]. The list of the used features and their dimensionality is shown in Table 1. We concatenate all features into one feature matrix, leading to 103 feature dimensions.

Instead of using all time frames of the low-level feature vectors during classification stage, for each of the songs in he database we calculate first four sample moments (namely *mean*, *standard deviation*, *skewness* and *kurtosis*) for each feature dimension. This results to a vector of length 412 for each of the songs.

3. FEATURE SELECTION

The aim of FS is to choose the minimal size set of informative features regarding class separation. The information provided by the chosen feature should be of minimal redundancy. According to the definitions proposed in [6] FS algorithms can be embedded in the classifier, can be used as a filter preceding to the classifier, or can use a learning algorithm as a subroutine. The approach applied here is of filter type.

3.1. Inertia Ratio Maximization using Feature Space Projection

IRMFSP was proposed by Peeters and Rodet [1]. This FS algorithm is motivated by the ideas similar to Fisher's discriminant analysis. On each iteration of the algorithm we look for the feature maximizing the ratio of between-class inertia to the total-class inertia. To avoid that on the next iteration the

Table 1: Low-level audio features used in the system

Feature	Short Name	Dimension
Log Loudness	LogLoud	12
Norm Loudness	NormLoud	12
Mel-frequency Cepstral Coefficients	MFCC	16
Audio Spectrum Envelope	ASE	14
Spectral Centroid	CENT	16
Spectral Crest Factor	SCF	16
Spectral Flatness Measure	SFM	16
Zero Crossing Rate	ZCR	1

next chosen feature could bring the same information, all features a orthogonalized to already selected one. The algorithm can be stopped, when the desired number of feature is chosen, or when the relative change of observed inertia ratio fulfill predefined conditions. In this evaluation we use the ISMFSP algorithms with the modifications proposed in [2].

4. FEATURE SPACE TRANSFORMATION

4.1. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is one of the most often used supervised dimension reduction methods [3], it is successfully applied as a preprocessing for audio signal classification. Original feature vectors are linearly mapped into new feature space guaranteeing a maximal linear separability by maximization of the ratio of between-class variance to the within-class variance. This mapping is conducted by multiplying the original $K \times N$ dimension feature matrix $\mathbf{X}$ with the transformation matrix $\mathbf{T}$. Reducing the dimension of the transformed feature vector from N to $D \leq N$ is achieved by considering only the first D column vectors of $\mathbf{T}$ for multiplication.

4.2. Generalized Discriminant Analysis

Real-world classification routines often have to deal with non-linear problems, thus linear discrimination in the original feature space is often not possible. General Discriminant Analysis (GDA) was firstly proposed in [7]. The idea of GDA is to map the features into higher dimensional (sometimes infinity dimensional) space, where the linear discrimination is possible. Dealing with a high dimensional space leads to an increase of the computation effort. To overcome this problem, the so called *kernel trick* is applied. The key idea of the kernel trick is to replace the dot product in a high-dimensional space with a kernel function in a original feature space.

5. CLASSIFIERS

In this section we shortly describe the applied classifiers. All in all the classifiers can be subdivided into two principal classes: namely *generative* and *discriminative* ones. Generative classifiers create models for the classes, describing the probability density distribution of the samples for each of the classes, and then use these models in the classification stage. Discriminative models do not try to model the distribution of the data samples and concentrate on determining the boundaries between the classes. Both generative and discriminative classifiers can either use rather sophisticated approximation routines (like Support Vector Machines or Gaussian Mixture Models) or in contrast apply rather straight forward approaches.

5.1. Support Vector Machines

A Support Vector Machine (SVM) is a discriminative classifier, attempting to generate an optimal decision plane between feature vectors of the training classes [8]. Commonly for real-world applications, the classification with linear separation planes is not possible in the original feature space. The transformation to the higher dimensional space is done using above mentioned kernel trick. Transformed into a high-dimensional space, non-linear classification problems can become linearly solvable.

5.2. Gaussian Mixture Models

Gaussian Mixture Models (GMM) is a well performing commonly used generative classifier. Single data samples of the class are thought of as generated from various sources and each source is modeled by a single multivariate Gaussian. The probability density function (PDF), i.e. the class model, is estimated as a weighted sum of the multivariate normal distributions. The parameters of the GMM can be estimated using the Expectation-Maximization algorithm [9].

5.3. k-Nearest Neighbor

With k-Nearest Neighbor (kNN), the classification is based on the class assignment of the closest training examples in the feature space [10]. This type of discriminative classifier is also referred as instance based learning. The level of generalization of kNN can be tuned by setting the number of nearest neighbors k taken into account.

5.4. Naive Bayes Classifier

Naive Bayes classifier (NB) is a simple probabilistic classifier with a possibility to model a priori probabilities of classes. NB uses a strong assumption of feature dimensions being statistically independent and thus takes into account only means and variances over the feature dimensions for all training data of the class. Recently, applicability and efficiency of NB classifiers have been discussed in [11] in detail.

6. EVALUATION

6.1. Database

For quantitative evaluation of the proposed FS and FST techniques, a test-set of full-length music pieces has been assembled. Altogether, the test-set consists of 775 tracks, belonging to 10 musical genres given in Table 2 (see [12] for details). The database is randomly subdivided into training (70%) and test (30%) set. All parameter tuning evaluations are run on the training part of the database.

Table 2: Music genres in the database

Music genre	Number of Songs
Classical	61
Electronic	170
Jazz	113
Pop	75
Rock	92
Ger. Pop	54
Urban	110
Speech	31
World	53
Misc.	16

6.2. Evaluation settings of algorithms

Here we provide a summary and some details of the tuning and parameter ranges of the applied algorithms.

Feature extraction: Some of the applied algorithms (e.g., GDA) require the features to be centered in the original feature space. Thus, we performed the normalization to zero mean and unit covariance. The normalization has been applied to the training set, and then the parameters of this normalization have been reused to normalize the test set in a same manner.

IRMFSP: Following the suggestion of [2] we fixed the number of selected feature dimensions. We varied the number of selected features from 20 to 200 with a step-width of 20 during the experiments.

LDA: As it was shown in theoretical works [3], for the classification of an L-class problem, at least $L-1$ feature dimensions are required. In this work the number of feature dimensions after LDA transform was set to $L-1$, where L is the amount of classes.

GDA: A Radial Basis Function (RBF) $K(x,y) = \exp -\gamma\|x-y\|^2$ was chosen for kernel transformation, with a kernel parameter γ with in a range of 2^{-15} and 2^{-5}. The optimal kernel parameter was chosen during 5-fold cross validation on the training set for each classifier independently. The feature vectors have been mapped into $L-1$ dimensions.

SVM: As wit the GDA algorithm, here we also used an RBF kernel. Kernel parameter γ was varied on logarithmic scale within a range of 2^{-15} and 2^{-5}. The search for the optimal cost parameter C was set within a range of 2^0 and 2^{10}, also on logarithmic scale. Originally, SVM is designed for binary problems. Here we used the "one-against-one" voting strategy, which requires training $(L(L-1)/2)$ binary classifiers to solve an L-class problem.

GMM: Here, we assumed the feature dimensions to be statistically independent and thus used diagonal covariance matrices. The initialization of the GMM was performed with the k-means algorithm, which was initialized randomly. In the work we estimated GMM with 2,3,4, and 10 Gaussian mixtures.

NB: We do not use the information about a priori probability of the music genres. Thus the Bayes formula is not explicitly used. We consider that all feature dimensions are statistically independent and restrict modeling of each class to a single multivariate normal distribution with a diagonal covariance matrix.

kNN: We do not restrict the number of samples to train the kNN classifier. I.e. all available training items are used for testing. The experiments have been conducted for $k=1$, $k=5$, $k=10$.

7. RESULTS

The classification result of the baseline system (without using FS and FST algorithms) is presented in Table 3. Results reflect the misclassification rate in percent. The best result of 31.69% are achieved when using SVM. The optimal parameter for the SVM classifier are estimated during 5-fold cross-validation on the training set. With GMM classifiers, the best result is achieved for a GMM with 2 mixtures. Satisfying classification rates also reached with NB classifier.

LDA is one of the most often used supervised dimension reduction algorithm. It it often applied to both linear solvable and linear non-solvable classification problems. In our work, linear transformation of the feature space to $L-1$ dimensions did not bring satisfying results. For all classifiers, the misclassification rates significantly increased (see Table 4).

In contrast to LDA, the non-linear feature space transformation definitely improved the results for all classifiers. While the classification rate for SVM and GMM changed only slightly, NB and kNN showed significant improvement (compare Table 5 and Table 3).

Using feature selection IRMFSP algorithm on the original features showed, that with a smaller set of feature dimensions one can receive comparable classification rates for SVM and GMM classifiers. For the optimal number of $100-120$ feature dimensions the results have been even improved. Choosing most informative feature dimensions significantly refined classification rate for NB and kNN classifiers. For the optimum number of selected dimensions they nearly graded up to SVM and GMM (compare Table 6 and Table 3). The best result for each classifier is marked with fold font.

Applying LDA transform after IRMFSP feature selection improved GMM, NB and kNN classifiers (see Table 7) while results for SVM classifier degraded a bit. Bold font marks the best classification rates. As one can see (compare Table 7 and Table 6) the best results are achieved approximately for the same number of selected features.

The best results for all classifiers have been achieved while utilizing GDA transform after IRMFSP feature selection (see Table 8). In this case the classification rates for all applied classifiers do not differ strongly. That is, even with rather simple classifiers like NB and kNN one can achieve the same performance as with SVM or GMM.

8. CONCLUSIONS

In MIR research, feature dimensionality reduction methods are often used blindly and lead to information loss. Thus, sometimes they not only do not improve classification results, but even impair the accuracy. In this paper we presented a comparative study on applying various feature selection and feature space transformation algorithms and tested the classification results with four different classifiers. We showed, that blind usage of supervised linear discrimination in probably non-linear feature spaces can decrease the classification rates. We demonstrated the power of utilizing a non-linear discrimination approach in MIR framework. Additionally we displayed that the well-suited FS and FST techniques can raise the performance of rather simple straight forward classifiers to the level of sophisticated ones.

Table 3: Classification results for data in original feature space, misclassification in %

Data	Dim.	SVM	GMM2	GMM3	GMM5	GMM10	NB	kNN1	kNN5	kNN10
Orig.	412	31.69	38.27	38.68	39.92	44.03	39.92	46.50	44.03	51.03

Table 4: Classification results for data after LDA transform, misclassification in %

Data	Dim.	SVM	GMM2	GMM3	GMM5	GMM10	NB	kNN1	kNN5	kNN10
LDA	9	60.08	90.12	91.36	90.95	90.53	67.49	60.49	60.08	60.08

Table 5: Classification results for data after GDA transform, misclassification in %

Data	Dim.	SVM	GMM2	GMM3	GMM5	GMM10	NB	kNN1	kNN5	kNN10
GDA	9	31.69	35.80	32.92	36.63	41.15	31.69	34.57	31.28	32.1

Table 6: Classification results for data after IRMSFP selection, misclassification in %

Data	Dim.	SVM	GMM2	GMM3	GMM5	GMM10	NB	kNN1	kNN5	kNN10
IRMFSP20	20	38.68	41.98	39.92	41.56	42.39	43.62	47.33	43.21	39.92
IRMFSP40	40	32.10	40.74	39.92	42.80	39.92	37.04	39.92	37.45	36.63
IRMFSP60	60	34.16	36.21	37.86	33.33	44.03	36.21	39.09	37.86	**34.16**
IRMFSP80	80	35.09	37.45	37.86	35.80	**37.45**	**35.80**	39.92	36.63	38.27
IRMFSP100	100	32.1	36.63	39.92	38.68	39.92	**35.80**	**37.86**	**34.98**	36.21
IRMFSP120	120	**29.63**	37.45	34.98	**34.57**	44.86	36.21	39.92	37.45	37.04
IRMFSP140	140	31.28	38.68	**34.57**	37.86	42.80	36.63	42.80	36.63	37.04
IRMFSP160	160	32.51	**33.74**	37.45	38.27	40.74	**35.80**	40.33	37.86	39.09
IRMFSP180	180	35.8	37.45	39.09	41.15	40.33	36.21	41.15	38.68	41.15
IRMFSP200	200	34.57	38.27	40.33	39.92	42.8	37.86	43.21	41.56	44.03

Table 7: Classification results while applying LDA after IRMFSP, misclassification in %

Data	D	SVM	GMM2	GMM3	GMM5	GMM10	NB	kNN1	kNN5	kNN10
IRMFSP20+LDA	9	41.98	39.92	41.98	44.86	44.86	41.56	46.09	44.86	43.62
IRMFSP40+LDA	9	34.57	36.21	34.98	36.63	38.27	37.45	38.68	35.80	35.80
IRMFSP60+LDA	9	35.80	34.98	35.80	39.51	38.27	37.04	41.56	36.63	37.45
IRMFSP80+LDA	9	32.92	34.57	36.21	39.51	**35.80**	34.16	38.68	33.74	**32.51**
IRMFSP100+LDA	9	33.33	32.92	37.04	37.04	38.68	32.92	39.92	**33.33**	32.92
IRMFSP120+LDA	9	**32.51**	**31.28**	**32.51**	**33.74**	39.09	**30.22**	35.39	33.74	32.92
IRMFSP140+LDA	9	33.33	31.69	34.98	39.92	38.68	31.69	**33.74**	34.57	34.16
IRMFSP160+LDA	9	34.98	39.92	38.68	42.80	43.62	36.21	38.68	36.63	35.39
IRMFSP180+LDA	9	35.80	44.44	45.27	44.44	48.15	34.98	37.86	34.57	35.39
IRMFSP200+LDA	9	34.57	44.86	45.27	49.38	47.74	37.04	39.09	36.63	36.63

Table 8: Classification results while applying GDA after IRMFSP, misclassification in %

Data	D	SVM	GMM2	GMM3	GMM5	GMM10	NB	kNN1	kNN5	kNN10
IRMFSP20+GDA	9	37.86	36.63	35.80	37.04	40.33	37.45	39.51	37.45	37.45
IRMFSP40+GDA	9	37.45	31.69	31.69	36.63	36.21	30.86	35.80	33.74	32.10
IRMFSP60+GDA	9	35.39	34.98	34.57	37.04	36.63	33.33	33.74	31.69	32.10
IRMFSP80+GDA	9	34.57	32.92	31.69	35.39	35.80	33.74	32.92	32.10	32.51
IRMFSP100+GDA	9	31.28	31.28	32.51	35.80	**34.57**	32.10	31.28	31.28	31.28
IRMFSP120+GDA	9	30.86	**28.81**	30.86	34.16	35.80	**29.63**	30.86	30.86	30.86
IRMFSP140+GDA	9	**28.81**	**28.81**	**28.81**	32.92	35.39	31.28	**28.81**	**28.81**	**28.81**
IRMFSP160+GDA	9	30.45	30.45	30.45	33.74	36.21	31.69	30.45	30.45	30.04
IRMFSP180+GDA	9	31.69	31.69	32.10	**32.10**	35.39	32.51	31.69	31.69	31.69
IRMFSP200+GDA	9	30.45	30.45	30.45	33.74	33.74	32.10	30.45	30.45	30.45

9. REFERENCES

[1] Geoffroy Peeters and Xavier Rodet. Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instruments databases. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx)*, London, UK, 2003.

[2] Slim Essid. *Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique*. PhD thesis, l'Université Pierre et Marie Curie, Paris, France, December 2005.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, September 1990.

[4] B.P. Bogert, M.J.R. Healy, and J.W. Tukey. The frequency analysis of time series for echoes: cepstrum, pseudoautocovariance, cross-cepstrum, and saphe cracking. In *Proceedings of the Symposium on Time Series Analysis*, 1963.

[5] H.-G. Kim, N. Moreau, and Th. Sikora. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Wiley & Sons, October 2005.

[6] L. C. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2002.

[7] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

[8] V.N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[9] A.P. Dempster, N. M. Laird, and D. B. Rdin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[10] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2nd edition, November 2000.

[11] H. Zhang. The optimality of naive bayes. In *Proceedings of the FLAIRS Conference*. AAAI Press, 2004.

[12] Christian Dittmar, Christoph Bastuck, and Matthias Gruhne. Novel mid-level audio features for music similarity. In *Proc. of the Intern. Conference on Music Communication Science (ICOMCS)*, Sydney, Australia, 2007.